DOCUMENT RESUME

ED 306 282 TM 013 139

AUTHOR Schumacker, Randall E.
TITLE Relationship between Multiple Regression and Selected Multivariable Methods.
PUB DATE Mar 89
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Discriminant Analysis; Equations (Mathematics); *Factor Analysis; Mathematical Models; *Multiple Regression Analysis; Multitrait Multimethod Techniques; *Multivariate Analysis; *Path Analysis; Regression (Statistics)
IDENTIFIERS Linear Models; *LISREL Computer Program

ABSTRACT

The relationship of multiple linear regression to various multivariate statistical techniques is discussed. The importance of the standardized partial regression coefficient (beta weight) in multiple linear regression as it is applied in path, factor, LISREL, and discriminant analyses is emphasized. The multivariate methods discussed in this paper have in common the general linear model and are the same in several other respects: (1) they identify, partition, and control variance; (2) they are based on linear combinations of variables; and (3) the linear weights can be computed based on standardized partial regression coefficients. However, these methods have different applications. While multiple regression seeks to identify and estimate the amount of variance in the dependent variable attributed to one or more independent variables, path analysis attempts to identify relationships among a set of variables. Factor analysis tries to identify subsets of variables from a much larger set. The LISREL program determines the degree of model specification and measurement error. Discriminant analysis seeks to identify a linear combination of variables that can be used to assign subjects to groups. An understanding of multiple regression and general linear model techniques can greatly facilitate one's understanding of the testing of research questions in multivariate situations. Eight appendices contain computer program examples based on correlational input as illustrations of these methods. A 47-item list of references is provided. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306282

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Randall E. Schumacker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Relationship between Multiple Regression and Selected Multivariable Methods

by

Randall E. Schumacker, Ph.D.
University of North Texas

American Educational Research Association
March 28, 1989

TM013139

BEST COPY AVAILABLE

Preface

The appropriate statistical method to use is often an issue of debate. It sometimes requires more than one approach to analyze data. The rationale for choosing between the alternative methods of analysis is usually guided by:

a. purpose of the research
b. research hypothesis or question
c. mathematical characteristics of the variables
d. sampling procedures
e. statistical assumptions
f. model validity

Multiple linear regression as a general linear model technique provides an excellent educational framework in which to analyze univariable and multivariable research questions (Newman, 1988). The present paper extends the relationship of multiple linear regression to various multivariable techniques: path, factor, LISREL, and discriminant analyses. The primary focus of which is to indicate the use of the standardized partial regression coefficient (beta weight) in these multivariable techniques.

This paper did not concern itself with issues of standardized versus unstandardized regression coefficients, Type I and Type II error rates, R-square shrinkage, suppressor variables, number of predictors, multicollinearity, curvilinearity and trend analysis, and many other issues related to general linear model research. Although, model specification and measurement error were addressed as an advantage of the LISREL approach.

Table of Contents

Topic	Page Number
Preface	i
Table of Contents	ii
Introduction	1
Multiple Regression	2
Multiple regression example	
Path Analysis	5
Model specification	
Path analysis example	
Factor Analysis	10
Factor models	
Factor analysis example	
LISREL	15
Measurement error	
LISREL-Factor analysis example	
LISREL-Regression analysis example	
Discriminant Analysis	19
Conclusion	24
Appendices	27
Appendix A - Multiple Regression Program	
Appendix B - Path Analysis Program(s)	
Appendix C - Factor Analysis Program	
Appendix D - LISREL Factor Analysis Program	
Appendix E - LISREL Regression Analysis Program	
Appendix F - Discriminant Analysis Program(s)	
Appendix G - Pascal Program	
Appendix H - SAS PC Program	
References	37

INTRODUCTION

Multiple Regression or the general linear model approach to the analysis of experimental data in educational research has become increasingly popular since 1967 (Bashaw and Findley, 1968). In fact today, it has become recognized as an approach that bridges the gap between correlational and analysis of variance thought in answering research hypotheses (McNeil, Kelly, & McNeil, 1975). Statistical textbooks in psychology and education often present the relationship between data analysis with multiple regression and analysis of variance (Draper & Smith, 1966; Williams, 1974; Roscoe, 1975; Edwards, 1979). Graduate students taking an advanced statistics course are therefore provided with the multiple linear regression framework for data analysis. Given their understanding of multiple linear regression techniques applied to univariate analysis (one dependent variable), their understanding can be extended to the relationship of multiple linear regression to various multivariate statistical techniques (Kelly, Beggs, McNeil, with Eichelberger & Lyon, 1969, pps 228-248). The present paper will expand upon this understanding and indicate the importance of the standardized partial regression coefficient (beta weight) in multiple linear regression as it is applied in path, factor, LISREL and discriminant analyses.

MULTIPLE REGRESSION

Multiple Regression techniques require a basic understanding of sample statistics (n, mean, and variance), standardized variables, correlation (Pedhazur, 1982, pp 53-57), and partial correlation (Cohen & Cohen, 1975; Houston & Bolding, 1974). In standard score form the multiple regression equation is:

$$\hat{z}_y = \beta z_x$$

The relationship between the correlation coefficient, the unstandardized regression coefficient and the standardized regression coefficient is:

$$\beta = \frac{\Sigma z_x z_y}{\Sigma z_x^2} = b \frac{s_x}{s_y} = r_{xy}$$

For two independent variables, the regression equation with standard scores is:

$$\hat{z}_y = \beta_1 z_1 + \beta_2 z_2$$

And the standardized partial regression coefficients are computed by:

$$\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \qquad \beta_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$$

The correlation between the original and predicted scores is given the special name *Multiple Correlation Coefficient*. It is indicated as:

$$R_{y\hat{y}} = R_{y.12}$$

And the *Squared Multiple Correlation Coefficient* is related as follows:

$$R^2_{y\hat{y}} = R^2_{y.12} = \beta_1 r_{y1} + \beta_2 r_{y2}$$

MULTIPLE REGRESSION EXAMPLE

A multiple linear regression example using a correlation matrix as input (SPSSX User's Guide, 3rd Edition, 1988, Chapter 13) is provided in Appendix A. The results are:

$$R^2_{y.123} = \beta_1 r_{y1} + \beta_2 r_{y2} + \beta_3 r_{y3}$$

$$= (.423)\ .507 + (.363)\ .481 + (.040)\ .276$$

$$R^2_{y.123} = .40$$

A systematic determination of the most important set of variables can be accomplished by setting the partial regression weight of each variable to zero. This approach and other alternative methods are presented by Kelly (1969) and Darlington (1968).

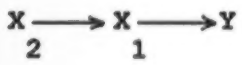
In summary, regression techniques have been shown to be robust (Bohrnstedt & Carter, 1971); applicable to contrast coding (Lewis & Mouw, 1978); dichotomous coding (McNeil, Kelly, & McNeil, 1975); and ordinal coding (Lyons, 1971) research situations. Multiple regression can also be viewed as a special case of path analysis.

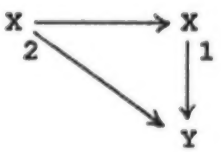
PATH ANALYSIS

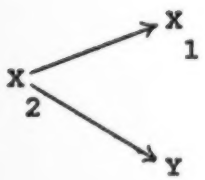
Sewall Wright is credited with the development of path analysis as a method for studying the direct and indirect effects of variables (Wright, 1921, 1934, 1960). Path analysis is not a method for discovering causes, rather a model must be specified by the researcher, similar to hypothesis testing in regression analysis. The specified model establishes causal relationships among the variables when:

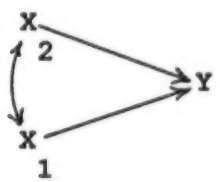
a. temporal ordering exists
b. covariation (correlation) is present
c. other causes controlled for

Model specification is necessary in examining multiple variable relationships. In the absence of a model, many different relationships among variables can be postulated with many different path coefficients being selected. For example, in a three variable model, the following relationships could be postulated:

(a) 

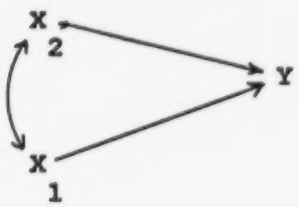
(b) 

(c) 

(d) 

The four different models have been considered without reversing the order of the variables. How can we decide which model is the correct one? Path analysis doesn't provide a way to specify the model, but rather estimates the effects once the model has been specified "a priori".

Path coefficients in path analysis take on the value of a product-moment correlation and/or standardized regression coefficients in a model (Wolfle, 1977). For example given model (d):



THEN:

$\beta_1 = p_{y1}$ $\beta_2 = p_{y2}$ $r_{12} = p_{12}$

A different set of terms is also used to describe the relationships among variables. The following terminology should help:

endogeneous - dependent variable

exogenous - independent variable

p - path coefficient

p_e - path coefficient error

--> - causal path

<--> - correlated path

A path model is specified by the researcher based on theory or prior research. Variable relationships once specified, in standard score form, become standardized regression coefficients. In multiple regression, a dependent variable is regressed in a single analysis on all the independent variables. In path analysis one or more multiple regression analyses are performed. Path coefficients are computed based upon only the particular set of independent variables that lead to the dependent variable under consideration. As in regression analysis, path analysis can handle dichotomous and ordinal data, but special coding and interpretation is necessary (Boyle, 1970; Lyons, 1971).

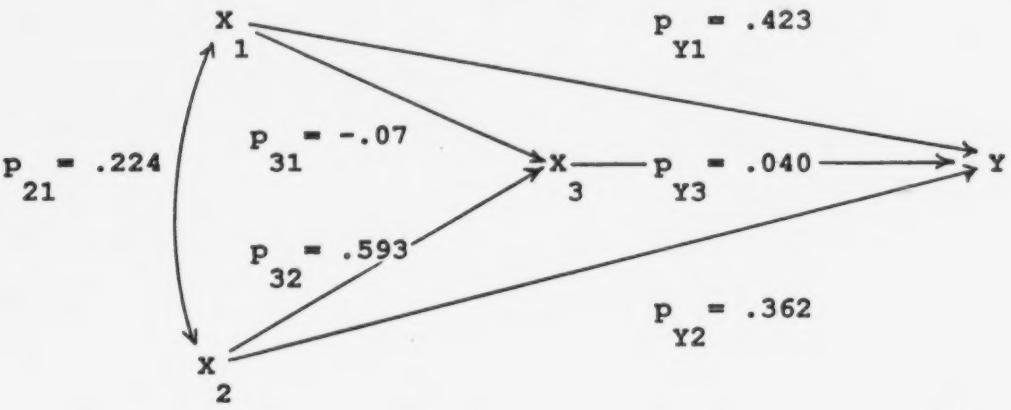
MODEL SPECIFICATION

Path models permit diagramming how a particular set of independent variables lead to a dependent variable under consideration. How the paths are drawn determine whether the independent variables are correlated causes (unanalyzed), mediated causes (indirect), or independent causes (direct). The model can be tested for the significance of path coefficients (Pedhazur, 1982, pp 58-62) and a goodness-of-fit criteria (Marascuilo & Levin, 1983, pp 169-172; Tatsuoka & Lohnes, 1988, pp 98-100) which reflects the significance between the original and reproduced correlation matrix. This process is commonly called

decomposing the correlation matrix (Asher, 1976, pp 32-34) according to certain rules (Wright, 1934).

PATH ANALYSIS EXAMPLE

A four variable model path analysis is presented in Appendix B. In order to calculate the path coefficients for the model, two regression analyses were performed. The model, with the path coefficients is:



The original and reproduced correlations are presented in matrix form. The upper half represents original correlations and the lower half the reproduced correlations which include only the regression of direct paths linking independent variables to the dependent variable.

VARIABLE	Y	X 1	X 2	X 3
Y	1.000	.507	.481	.276
X 1	.423	1.000	.224	.062
X 2	.362	.224	1.000	.577
X 3	.040	-.070	.593	1.000

Original Correlations

Reproduced Correlations

The original correlations can be completely "reproduced" if all effects: direct (DE), indirect (IE), spurious (S) and correlated (C) are included. For example:

$$r_{12} = \underset{C}{p_{12}} = .224$$

$$r_{13} = \underset{DE}{p_{31}} + \underset{IE}{p_{32}\,p_{21}} = .062$$

$$r_{23} = \underset{DE}{p_{32}} + \underset{S}{p_{31}\,p_{21}} = .577$$

$$r_{1Y} = \underset{DE}{p_{Y1}} + \underset{IE}{p_{Y2}\,p_{21}} + \underset{IE}{p_{Y3}\,p_{31}} + \underset{IE}{p_{Y3}\,p_{32}\,p_{21}} = .507$$

$$r_{2Y} = \underset{DE}{p_{Y2}} + \underset{IE}{p_{Y3}\,p_{32}} + \underset{S}{p_{Y1}\,p_{21}} + \underset{S}{p_{Y3}\,p_{31}\,p_{21}} = .481$$

$$r_{3Y} = \underset{DE}{p_{Y3}} + \underset{S}{p_{Y1}\,p_{31}} + \underset{S}{p_{Y2}\,p_{32}} + \underset{S}{p_{Y1}\,p_{21}\,p_{32}} + \underset{S}{p_{Y2}\,p_{21}\,p_{31}} = .276$$

In summary, path analysis can be carried out within the context of ordinary regression analysis and does not require the learning of any new analysis techniques (Asher, 1976, p32; Williams, 1974). The advantage of path analysis is that it enables one to specify direct and indirect effects among independent variables. In addition, path analysis enables us to decompose the correlation between any two variables into simple and complex paths of which some are meaningful. Path coefficients and the relationship between the original and reproduced correlation matirx can also be tested for significance.

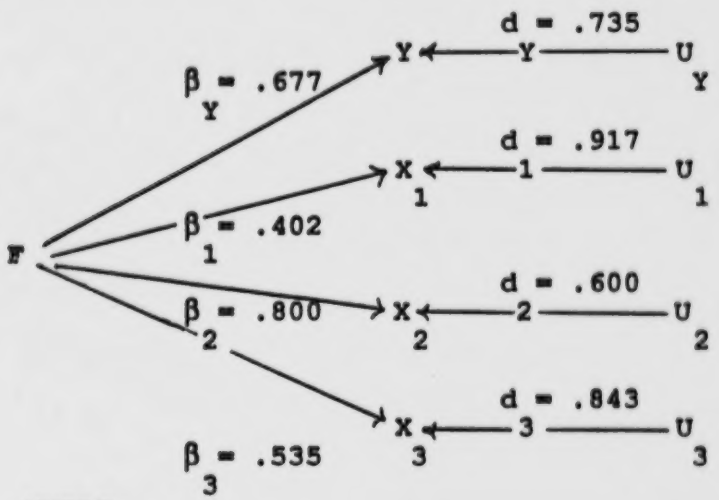
FACTOR ANALYSIS

Path models and the associated test of significance between original and reproduced correlations are used in confirmatory factor analysis. Factor analysis assumes that the observed (measured) variables are linear combinations of some underlying source variable (factor). In practice, one estimates population parameters of the measured variables from a sample (with the uncertainties of model specification and measurement error). A linear combination of weighted variables relates to multiple regression in a single factor model and to a linear causal system (path analysis - "multiple" multiple regressions) in multiple factor models. Path diagrams therefore permit representation of the causal relationships among factors and observed (measured) variables in factor analysis.

In general, the first step in factor analysis involves the study of interrelationships among variables in the correlation matrix. Factor analysis will address the question of whether these subsets can be identified by one or more factors (hypothetical constructs). Confirmatory factor analysis is used to test specific hypotheses regarding which variables correlate with which constructs.

FACTOR MODELS

Factor analysis assumes that some factors, which are smaller in number than the number of observed variables, are responsible for the covariation among the observed variables. For example, given a unidimensional trait in a single factor model with four variables the diagram would be (Kim & Mueller, 1978a, p 35):



WHERE:

β_i = standardized regression coefficient; path coefficient; or common factor loading

d_i = residual coefficient; path error coefficient; or unique factor loading

The variance of each observed variable is therefore comprised of the proportion of variance determined by the common factor and the proportion determined by the unique factor, which together equal the total variance of each observed variable. Therefore:

$$s_i^2 = \beta_i^2 + d_i^2 = 1$$

The correlation between a common factor and a variable is:

$$r_{F,X_i} = \beta_i$$

The correlation between a unique factor and a variable is:

$$r_{U,X_i} = d_i$$

The correlation between observed (measured) variables sharing a common factor is:

$$r_{X_i,X_j} = \beta_i \beta_j$$

And finally, the variance attributed to the factor as a result of the linear combination of variables is:

$$h^2 = \frac{\Sigma \beta_i^2}{M} = R^2_{F.1234}$$

Where: M = number of variables

β_i^2 = squared factor loadings

Note: $\Sigma \beta_i^2$ = eigenvalue

β_i^2 = communality

FACTOR ANALYSIS EXAMPLE

A single factor model analysis with four variables in a correlation matrix format is in Appendix C. The path diagram is the same as above (Kim & Mueller, 1978, p 35) with the weights as follows:

$$\beta_Y = .677 \quad \beta_1 = .402 \quad \beta_2 = .800 \quad \beta_3 = .535$$

And, factor scores computed as:

$$F = \beta_y Y + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Multiplying the coefficients between pairs of variables gives the following correlation matrix:

CORRELATION MATRIX

VARIABLE	Y	X_1	X_2	X_3
Y	β_1^2	.27	.54	.36
1	.27	β_2^2	.32	.22
2	.54	.32	β_3^2	.43
3	.36	.22	.43	β_4^2

The common factor variance is:

$$R^2_{F.1234} = \frac{\Sigma \beta_i^2}{M} = \frac{.46 + .16 + .64 + .29}{4} = .39$$

The unique factor variance is:

$$1 - R^2_{F.1234} = \frac{\Sigma (1 - \beta_i^2)}{M} = \frac{.54 + .84 + .36 + .71}{4} = .61$$

In summary, factor loadings (variable weights) are standardized regression coefficients. As such, linear weighted combinations of variables loading on a factor are used to compute factor scores. The weights are also the correlation between the observed (measured) variables and the factor (hypothetical construct). If the variable correlations (weights) are squared and summed, they describe the proportion of variance determined by the common factor. This is traditionally known as the coefficient of determination, but termed communality in factor analysis. When all variables are standardized, then the linear weights are called standardized regression coefficients (regression analysis), path coefficients (path analysis), or factor loadings (factor analysis). The factor analysis approach is distinquished from regression or path analysis in that observed variable correlation is explained by a common factor (hypothetical construct). In factor analysis therefore the correlation between observed variables is the result of sharing a common factor rather than a variable being the direct cause (path analysis) or predictor of another (regression analysis).

LISREL

Linear structural relationships (LISREL) are often diagrammed by using multiple factor path models where the factors (hypothetical contructs) are viewed as latent traits (Joreskog & Sorbom, 1986, pp I.5-I.7). The LISREL model consists of two parts: the measurement model and the structural equation model. The measurement model specifies how the latent variables or hypothetical constructs are measured in terms of the observed (measured) variables and describes their measurement properties (reliability and validity). The structural equation model specifies the causal relationship among the latent variables and is used to describe the causal effects and the amount of unexplained variance. The LISREL model includes or encompasses a wide range of models, for example; univariate or multivariate regression models, confirmatory factor analysis, and path analysis models (Joreskog & Sorbom, 1986, pp I.3, I.9-I.12). Cuttance (1983) presents an overview of several LISREL submodels with diagrams and explanations. Wolfle (1982) presents an indepth presentation of a single model to introduce and clarify LISREL analysis. The LISREL program therefore permits regression, path, and factor analysis whereby model specification and measurement error can be assessed.

MEASUREMENT ERROR

Fuller (1987) extensively covers LISREL and factor analysis models and especially extends regression analysis to the case where the variables are measured with error. Wolfe (1979, pp 48-51) presents the relationship between LISREL, regression and path analysis especially in regards to how measurement error effects the regression coefficient (path coefficient). Errors of measurement in statistics has been studied extensively (Wolfe, 1979). Cochran (1968) studied it from four different aspects: (a) types of mathematical models, (b) standard techniques of analysis which take into account measurement error, (c) effect of errors of measurement in producing bias and reduced precision and what remedial procedures are available, and (d) techniques for studying error of measurement. Cochran (1970) also studied the effects of error of measurement on the squared multiple correlation coefficient.

LISREL-FACTOR ANALYSIS EXAMPLE

A LISREL factor analysis model program with a correlation matrix as input is given in Appendix D. The factor analytic model in matrix notation is:

$$X = \Lambda_x \xi + \theta_\delta$$

Where: X = observed variables
Λ = structural weights (factor loadings)
ξ = latent trait (factor)
θ_δ = error variance (unique variance)

The LISREL results are:

a. Λ = LAMBDA X (structural weights-factor loadings)

$Y = .677 \quad X_1 = .402 \quad X_2 = .800 \quad X_3 = .535$

b. θ_δ = THETA DELTA (unique factor variance)

$Y = .54 \quad X_1 = .84 \quad X_2 = .36 \quad X_3 = .71$

c. β^2 = LAMBDA X^2 (common factor variance)

$Y = .46 \quad X_1 = .16 \quad X_2 = .64 \quad X_3 = .29$

The concept of model specification and goodness of fit pertains to the original correlation matrix and the estimated correlation matrix. The estimated correlation matrix is:

$$\Sigma = \begin{matrix} .272 & & \\ .542 & .321 & \\ .362 & .215 & .427 \end{matrix}$$

The original correlation matrix is:

$$S = \begin{matrix} .507 & & \\ .481 & .224 & \\ .276 & .062 & .577 \end{matrix}$$

The Goodness of Fit Index (GFI) using the unweighted least squares approach (ULS) is then computed as:

$$GFI = 1 - 1/2 \text{ trace } (S - \Sigma)^2$$

$$GFI = 1 - 1/2 \ (1.308 - 1.02)^2$$

$$GFI = 1 - .041$$

$$GFI = .959$$

LISREL-REGRESSION ANALYSIS EXAMPLE

A LISREL regression model program with a correlation matirx as input is given in Appendix E. The regression model in matrix notation is:

$$Y = \Gamma X + \zeta$$

Where: Y = dependent variable
Γ = gamma matrix (beta weights)
X = independent variables
ζ = errors of prediction (error variance)

The LISREL results are the same as in the previous regression program:

$$R^2_{y.123} = \Gamma_1 r_{y1} + \Gamma_2 r_{y2} + \Gamma_3 r_{y3}$$

$$R^2_{y.123} = (.423)\ .507 + (.363)\ .481 + (.040)\ .276$$

$$R^2_{y.123} = .40$$

DISCRIMINANT ANALYSIS

The general approach in both two group and multiple-group discriminant classification is to construct a linear combination of variables which optimally classifies or assigns subjects to known groups (Huberty, 1974; 1975). In the two group dependent variable case where only one discriminant function is needed, regression and discriminant analysis are the same (Kerlinger & Pedhazur, 1973, pp 336-340; Thayer, 1986). They are compared and presented in Appendix F. They differ in the multiple group case where more than one discriminant function is computed.

The linear combination of weighted variables can be expressed as:

$$L = \beta_i X_i \ldots\ldots + \beta_n X_n$$

with ß values chosen to provide maximum discrimination between two populations. The ß's are constructed as linear combinations of the differences between variable means in the two groups:

$$d_j = \bar{X}_{1j} - \bar{X}_{2j}$$

WHERE:

VARIABLE	GROUP MEANS 1	0	d
X_1	5.2	3.6	1.6
X_2	2.8	2.4	.4

The pooled sample covariance is represented as:

$$s_{jq} = \frac{\sum_{i=1}^{2} \sum_{k=1}^{n_j} (X_{ijk} - \bar{X}_{ij})(X_{iqk} - \bar{X}_{iq})}{n_1 + n_2 - 2}$$

To provide maximum discrimination, the variation of values in L between the two groups should be greater than the variation in the values of L within the two groups. In fact, this is just the case:

$$SS_B = \sum_{i=1}^{2} n_i (\bar{L}_i - \bar{L})^2 \qquad SS_W = \sum_{i=1}^{2} \sum_{k=1}^{n_j} (L_{ik} - \bar{L}_i)^2$$

The ratio of these two can be thought of as a measure of the discriminatory power of L, in the sense that the larger the value of sums of squares between, to sums of squares within, the more L is reflecting between population variance as opposed to within population variance.

The multiple regression and discriminant statistics are therefore related as:

$$F = \frac{SS_B / df_1}{SS_W / df_2} = \frac{SS_{reg} / p}{SS_{error} / n_1 + n_2 - p - 1} = \frac{1.194/2}{1.306/7}$$

$$= \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)p} * D^2 = 3.20$$

AND:

$$R^2 = \frac{SS_{reg}}{SS_{total}} = c * D^2 = .1632\ (2.928) = .4778$$

(note: $SS_{total} = Npq = 2.5$)

The Mahalanobis D-squared and constant value are computed as:

$$D^2 = \sum_{j=1}^{p} \sum_{q=1}^{p} d_j\, d_q\, s^{jq}$$

$$= \frac{(n_1 + n_2)\,(n_1 + n_2 - 2)}{n_1\, n_2} * \frac{R^2}{1 - R^2} = 2.928$$

$$c = \frac{\frac{n_1\, n_2}{n_1 + n_2}}{(n_1 + n_2 - 2)} + \frac{\frac{n_1\, n_2}{n_1 + n_2} * D^2}{} = \frac{2.5}{15.319} = .1632$$

Regression weights are compared to discriminant weights (β = regression weights; b=discriminant weights) as:

$$b_i = \frac{\beta_i}{c}$$

THEN:

$$b_1 = \frac{.1389}{.1632} = .8511 \qquad b_2 = \frac{.1204}{.1632} = .7375$$

AND:

$$b_0 = -.5\,(\bar{L}_1 + \bar{L}_2) = -.5(7.080 + 4.153) = -5.617$$

WHERE:

$$\bar{L}_1 = b_{1\,1}\bar{X}_{11} + b_{2\,1}\bar{X}_{10} = .8511(5.2) + .7375(3.6) = 7.081$$

$$\bar{L}_2 = b_{1\,2}\bar{X}_{21} + b_{2\,2}\bar{X}_{20} = .8510(2.8) + .7375(2.4) = 4.153$$

The formulae indicate that the regression procedure can be used to produce a linear combination of weights which only differ by a constant value c (the choice of coding values for the dependent variable will change the value of c). The quantity D-squared is called Mahalanobis D-squared and it represents a measure of the distance between two means.

An extension to the multiple group discriminant case using eigenvalues also relates the sums of squares approach to several multivariate statistics (Marascuilo & Levin, 1983, Chapters 7 and 8). For the two group case ($\lambda = .91489$):

a. Roy's criterion

$$\theta = \frac{\lambda}{1 + \lambda} = \frac{SS_B}{SS_T} = R^2 = .4778$$

b. Fisher's F

$$F = \frac{N - p - 1}{p} \cdot \frac{\theta}{1 - \theta} = 3.20$$

c. Hotelling's T^2

$$T^2 = \frac{(N - 2)\,\theta}{1 - \theta} = 7.32$$

Cautionary Remark

Mueller & Cozad (1988) discuss standardization procedures used in SPSSX, BMDP, and SAS to determine standardized discriminant coefficients. They indicated that the within-group variance (SPSSX, BMDP) should be used rather than the total variance (SAS) because it removes between-group differences from the estimate. Moreover, because standardized weights are computed differently it causes erroneous interpretations of results (SPSSX and BMDP use the diagonal elements of the within-group covariance matrix; SAS uses the diagonal elements of the total covariance matrix). These major "canned" statistical programs have inconsistencies between them and also within them.

CONCLUSION

The appropriate statistical method to use is often an issue of debate. It sometimes requires more than one approach to analyze data. The rationale for choosing between the alternative methods of analysis is usually guided by:

a. purpose of the research
b. research hypothesis or question
c. mathematical characteristics of the variables
d. sampling procedures
e. statistical assumptions
f. model validity

The multivariable methods discussed in this paper have in common the general linear model and are the same in several respects. First, they identify, partition, and control variance. Second, they are based on linear combinations of variables. And third, the linear weights can be computed based on standardized partial regression coefficients.

The multivariable methods however have different applications. Multiple regression seeks to identify and estimate the amount of variance in the dependent variable attributed to one or more independent variables (prediction). Path analysis seeks to identify relationships among a set of variables (explanation). Factor anlaysis seeks to idenfify subsets of variables from a much larger set (common/shared variance). LISREL determines the degree of model specification and measurement error. Discriminant analysis seeks to identify a linear combination of variables

which can be used to assign subjects to groups (classification). The different methods were derived because of the need for prediction, explanation, common variance, model and measurement error assessment, and classification type applications.

Multiple Regression techniques are robust except for model specification and measurement errors (Borhnstedt, 1971). Multiple regression techniques are useful in understanding path, factor, LISREL, and discriminant applications. LISREL permits regression, path, and factor analyses whereby model specification and measurement error can be assessed. LISREL also permits univariate or multivariate least squares analysis in either single sample or multiple sample (across populations) research settings. An understanding of multiple regression and general linear model techniques can therefore greatly facilitate one's understanding of the testing of research questions in multivariable situations.

Multiple linear regression is also related to canonical correlation analysis, under which all parametric tests are subsumed as special cases (Knapp, 1978; Marascuilo & Levin, 1983). A recent presentation suggested that multivariate analyses are really univariate analyses and further illustrates that an understanding of multiple regression facilitates an understanding of multivariable methods (Newman, 1988). Some authors have presented multivariate analysis of variance using multiple regression methods

(Woodward & Overall, 1975), while other authors (Huberty & Morris, 1987) present an argument for a truly multivariate analysis.

As a final comment, it is well known that the correlation matrix has a central role in the analysis of multivariable data. In fact, it was used in the numerous computer program examples which assumed standardized variables. The inverse of the correlation matrix, however, also has important interpretations in multiple regression, factor and discriminant analyses (Raveh, 1985). Two main roles are: (a) near a diagonal matrix as p, the number of variables, increases in order for factor analysis to be meaningful; and (b) the estimated coefficients in multiple regression and discriminant analysis are obtained from the inverse matrix and thus conditioned on p specific variables.

APPENDICES

The following appendices contain computer program examples based upon correlational input (SPSSX User's Guide, 3rd ed., Chapter 13, 1988) with the exception of discriminant analysis. These programs were run on a mainframe computer (with some modification they can also run on the personal computer version). A PASCAL program was written and compiled to generate random variables (Borland, 1988), and is in APPENDIX G. The random variables were then correlated using a SAS PC program (SAS, 1988) in APPENDIX H. Although random data were generated, the relationships and principles presented in this paper also apply to research data.

APPENDIX A

MULTIPLE REGRESSION ANALYSIS PROGRAM

```
TITLE REGRESSION WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507   1.000
 .481    .224   1.000
 .276    .062    .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=Y/
           ENTER X1 X2 X3/
FINISH
```

APPENDIX B

PATH ANALYSIS PROGRAM

A. VARIABLE 3 REGRESSED ON VARIABLES 1 AND 2

```
TITLE PATH ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT  VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=X3/
           ENTER X1 X2/
FINISH
```

B. VARIABLE Y REGRESSED ON VARIABLES 1, 2, AND 3

```
TITLE PATH ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
REGRESSION MATRIX=IN(*)/
           MISSING=LISTWISE/
           VARIABLES=Y X1 X2 X3/
           DEPENDENT=Y/
           ENTER X1 X2 X3/
FINISH
```

APPENDIX C

FACTOR ANALYSIS PROGRAM

```
TITLE FACTOR ANALYSIS EXAMPLE WITH CORRELATION MATRIX INPUT
COMMENT VARIABLE MEANS=0; VARIANCES=1; CONSTANT=0
MATRIX DATA VARIABLES=Y X1 X2 X3/N=100
BEGIN DATA
1.000
 .507  1.000
 .481   .224  1.000
 .276   .062   .577  1.000
END DATA
FACTOR VARIABLES=Y X1 X2 X3/
       MATRIX=IN(COR=*)/
       CRITERIA=FACTORS(1)/
       EXTRACTION=ULS/
       ROTATION=NOROTATE/
       PRINT CORRELATION DET INITIAL EXTRACTION ROTATION/
       FORMAT SORT/
       PLOT=EIGEN/
FINISH
```

APPENDIX D

LISREL ANALYSIS PROGRAM

```
TITLE 'LISREL FACTOR ANALYSIS WITH CORRELATION MATRIX INPUT'
INPUT PROGRAM
NUMERIC DUMMY
END FILE
END INPUT PROGRAM
USERPROC NAME=LISREL
DATA FOR GROUP ONE
DA NG=1 NI=4 NO=100
LA
'Y' 'X1' 'X2' 'X3'
KM SY
1.000
 .507 1.000
 .481  .224 1.000
 .276  .062  .577 1.000
MO NX=4 NK=1 TD=DI,FR PH=ST
LK
'FACTOR'
PA LX
4 * 1
OU ULS SE TV PC RS VA FS SS MI
END USER
```

APPENDIX E

LISREL REGRESSION ANALYSIS PROGRAM

```
TITLE 'LISREL REGRESSION ANALYSIS WITH CORRELATION MATIRX'
INPUT PROGRAM
NUMERIC DUMMY
END FILE
END INPUT PROGRAM
USERPROC NAME=LISREL
DATA FOR GROUP ONE
DA NG=1 NI=4 NO=100
LA
'Y' 'X1' 'X2' 'X3'
KM SY
1.000
 .507 1.000
 .481  .224 1.000
 .276  .062  .577 1.000
MO NY=1 NX=3 PS=DI
OU ULS SE TV PC RS VA SS MI TO
END USER
```

APPENDIX F

DISCRIMINANT ANALYSIS

A. DISCRIMINANT ANALYSIS VIA REGRESSION PROGRAM

```
TITLE REGRESSION ANALYSIS WITH DICHOTOMOUS DEPENDENT
DATA LIST  RECORDS=1 /1 Y 1 X1 3 X2 5
BEGIN DATA
1 8 3
1 7 4
1 5 5
1 3 4
1 3 2
0 4 2
0 3 1
0 3 2
0 2 2
0 2 5
REGRESSION VARIABLES= Y X1 X2/
            DEPENDENT=Y/
            ENTER X1 X2/
            SAVE PRED (PSCORE)/
PRINT /1 Y X1 X2 PSCORE
EXECUTE
FINISH
```

B. DISCRIMINANT ANALYSIS VIA DISCRIMINANT PROGRAM

```
TITLE DISCRIMINANT ANALYSIS WITH DICHOTOMOUS DEPENDENT
DATA LIST  RECORDS=1/1 Y 1 X1 3 X2 5
BEGIN DATA
1 8 3
1 7 4
1 5 5
1 3 4
1 3 2
0 4 2
0 3 1
0 3 2
0 2 2
0 2 5
DISCRIMINANT GROUPS= Y(0,1)/VARIABLES=X1 X2/ANALYSIS=X1 X2/
            METHOD=DIRECT/SAVE=CLASS=PRDY/
            STATISTICS 11 12 13/
COMPUTE YHAT= -5.617 + .8510 * X1 + .7375 * X2
PRINT /1 Y X1 X2 PRDY YHAT
EXECUTE
FINISH
```

APPENDIX G

PASCAL PROGRAM [a]

```
program ran;
 const
          ns=100;
         rxy=0.5;
          a1=2.505922;
          a3=-15.73223;
          a5=23.54337;
          b2=-7.337743;
          b4=14.97266;
          b6=-6.016088;

var
   a11,a1sq,x1,x2,x3,y1,y2,y3,z1,br,r : real;
   ix : longint;
   out1 : text;
   k:integer;
procedure gauss ( var pgbr:real;var  pgix : longint);
     var q,pgr:real;
          pgiy:longint;

procedure randu (prix:longint; var priy:longint;var
yfl:real);
     begin
          priy:=0;yfl:=0;
          priy:=prix*65539;
          if priy<0 then
                         priy:=priy+214748364?+1;
          yfl:=priy;
          yfl:=yfl*0.4656613e-9;
     end;
  (see J. A. Byars & J. T. Roscoe for algorithm explanation)

   begin
      pgiy:=0; pgr:=0;
      randu(pgix,pgiy,pgr);
      pgix:=pgiy;
      pgr:=pgr-0.5;
      q:=pgr*pgr;
      pgbr:=((a1+(a3+a5*q)*q)*pgr)/(1+(b2+(b4+b6*q)*q)*q);
   end;
```

[a] Special acknowledgement to Miguel Monsivais, a doctoral student, in educational research who converted a prior FORTRAN program into Pascal code.

APPENDIX G (CONTINUED)

```
  { see T. Knapp & V. Swoyer for algorithm explanation }

begin
  ix:=0;
  assign(out1,'corr.dat');
  rewrite(out1);
  ix:=16875423;
alsq:=rxy*rxy;
 a11:=sqrt(1-alsq);
for k:=1 to ns do
   begin
      x1:=0;y1:=0;z1:=0;br:=0;
      gauss (br,ix);
      x1:=br;
      br:=0;
      gauss (br,ix);
      z1:=br;
      y1:=(rxy*x1)+(a11*z1);

      gauss(br,ix);
      x2:=br;
      y2:=(rxy*y1)+(a11*x2);

      gauss(br,ix);
      x3:=br;
      y3:=(rxy*y2)+(a11*x3);
      writeln (out1,y1:10:6,x1:10:6,y2:10:6,y3:10:6);
   end;
close(out1);
end.
```

APPENDIX H

A. SAS PC CORRELATION PROGRAM

```
data a;
infile 'c:corr.dat';
input y x1 x2 x3 @@;
proc corr;var y x1 x2 x3;
run;
```

References

Asher, H.B. (1976). *Causal modeling*. Sage Publications: Beverly Hills: CA.

Bashaw, W.L. & W.G. Findley (1968). *Symposium on general linear model approach to the analysis of experimental data in educational research*. Project No. 7-8096. U.S. Department of Health, Education, and Welfare, Washington, D.C.

Bohrnstedt, G.W. & T.M. Carter (1971). *Robustness in regression analysis*. In H.L. Costner (Ed), *Sociological Methodology*, Jossey-Bass, pp 118-146.

Turbo Pascal 4.0 (1988). Borland International, Scotts Valley: CA.

Boyle, R.P. (1970). Path analysis and ordinal data. *American Journal of Sociology*, vol. 75(4), 461-480.

Byars, J.A. & J.T. Roscoe (1972). *Rational approximations of the inverse gaussian function*. Presented at the American Educational Research Association Annual Convention, Chicago, IL.

Cohen, J. & P. Cohen (1975). *Applied multiple regression/ correlation analysis for the behavioral sciences*. Lawrence Erlbaum: NJ.

Cochran, W.G. (1968). Errors of measurement in statistics. *Technometrics*, 10(4), 637-666.

Cochran, W.G. (1970). Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, 65(329), 22-34.

Cuttance, P.F. (1983). Covariance structure and structural equation modelling in research: a conceptual overview of LISREL modelling. *Multiple Linear Regression Viewpoints*, 12(2), 1-63.

Darlington, R.B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.

Draper, N.R. & H. Smith (1966). *Applied regression analysis*. John Wiley & Sons: New York, NY.

Edwards, A.L. (1979). *Multiple regression and the analysis of variance and covariance*. W.H. Freeman: San Francisco, CA.

Fuller, W.A. (1987) *Measurement error models*. John Wiley & Sons: New York, NY.

Glass, G.V. & K.D. Hopkins (1984). *Statistical methods in education and psychology*. Prentice-Hall: Englewood Cliffs, NJ.

Houston, S.R. & J.T. Bolding, Jr. (1974). *Part, partial, and multiple correlation in commonality analysis of multiple regression models. Multiple Linear Regression Viewpoints*, 5(3), 36-40.

Huberty, C.J. & J.D. Morris (1987). Multivariate analysis versus multiple univariate analyses. *Multiple Linear Regression Viewpoints*, 16(1), 108-125.

Huberty, C.J. (1974). *Discriminant analysis*. Presented at American Educational Research Association, Chicago, IL.

Huberty, C.J. (1975). Discriminant analysis. *Review of Educational Research*, 45(4), 543-598.

Joreskog, K.G. & D. Sorbom (1986). *LISREL VI USER'S GUIDE: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Scientific Software: Mooresville, Indiana.

Kelly, F.J, D.L. Beggs, K.A. McNeil, T. Eichelberger, & L. Lyon (1969). *Multiple regression approach*. Southern Illinois University Press: Carbondale, IL.

Kerlinger, F.N. & E.J. Pedhazur (1973). *Multiple regression in behavioral research*. Holt, Rinehart & Winston, New York: NY.

Kim, J. & C. Mueller (1978a). *Introduction to Factor Analysis*. Sage Publications, Beverly Hills: CA.

Kim, J. & C. Mueller (1978b). *Factor Analysis*. Sage Publications, Beverly Hills: CA.

Knapp, T.R. & V.H. Swoyer (1967). Some empirical results concerning the power of Bartlett's test of the significance of a correlation matrix. *American Educational Research Association*, 4(1), 13-17.

Knapp, T.R. (1978). Canonical correlation analysis: a general parametric significance-testing system. *Psychological Bulletin*, 85(2), 410-416.

Lewis, E.L. & J.T. Mouw (1978). *The use of contrast coefficients*. Southern Illinois University Press, Carbondale, IL.

Lyons, M. (1971). *Techniques for using ordinal measures in regression and path analysis.* In H.L. Costner (Ed), *Sociological Methodology*, Jossey-Bass, pp 147-171.

Marascuilo, L.A. & J.R. Levin (1983). *Multivariate statistics in the social sciences: a researcher's guide.* Brooks/Cole Publishing: Belmont: CA.

McNeil, K.A., F.J. Kelly, J.T. McNeil (1975). *Testing research hypotheses using multiple linear regression.* Southern Illinois University Press: Carbondale, IL.

Newman, I. (1988). *There is no such thing as multivariate analysis: all analyses are univariate.* Presidential Address at Midwestern Education Research Association, Chicago, IL.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research: explanation and prediction (2nd ed.).* Holt, Rinehart, & Winston: New York, NY.

Raveh, A. (1985). On the use of the inverse of the correlation matrix in multivariate data analysis. *American Statistical Association*, 39(1), 39-42.

Roscoe, J.T. (1975). *Fundamental research statistics for the behavioral sciences (2nd ed.).* Holt, Rinehart, & Winston: New York, NY.

SAS User's Guide, 6.03 (1988). SAS Institute Inc, Cary, NC.

SPSSX Users' Guide, 3rd ed. (1988). McGraw-Hill: New York, NY.

Tatsuoka, M.M & P.R. Lohnes (1988). *Multivariate analysis: techniques for educational and psychological research (2nd ed.).* Macmillan Publishing Company: New York, NY.

Thayer, J. (1986). Using multiple regression with dichotomous dependent variables. *Multiple Linear Regression Viewpoints*, 15(1), 90-98.

Williams, J.D. (1974). *Regression analysis in educational research.* MSS Information Corporation: New York, NY.

Williams, J.D. (1974). Path analysis and causal models as regression techniques. *Multiple Linear Regression Viewpoints*, 5(3), 1-20.

Wolfle, L.M. (1977). An introduction to path analysis. *Multiple Linear Regression Viewpoints*, 8(1), 36-61.

Wolfle, L.M. (1979). Unmeasured variables in path analysis. *Multiple Linear Regression Viewpoints*, 9(5), 20-56.

Wolfle, L.M. (1982). Causal models with unmeasured variables: an introduction to LISREL. *Multiple Linear Regression Viewpoints*, 11(2), 9-54.

Woodward, J.A. & J.E. Overall (1975). Multivariate analysis of variance by multiple regression methods. *Psychological Bulletin*, 82, 21-32.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161-215.

Wright, S. (1960). Path coefficients and path regression: alternative or complementary concepts? *Biometrics*, 16, 189-202.